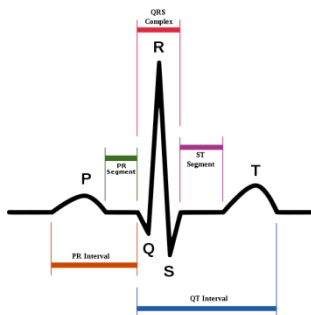


Zestawy danych

heart_c.arff

Baza danych opisująca wyniki badań pacjentów mających problemy z sercem. Oryginalna baza danych posiadała 76 atrybutów. Tylko 14 wykorzystano w eksperymentach naukowych i tylko na tych 14 będziemy pracować. Każdy wiersz tabelki dotyczy jednego pacjenta

- Age: wiek pacjenta w latach – liczba rzeczywista
- Sex: płeć pacjenta – atrybut nominalny, dwie wartości: male, female
- Cp: typ bólu w piersi - atrybut nominalny, cztery wartości:
 - typ_angina – typowy ból wynikający ze złego dostarczania tlenu do mięśnia sercowego.
 - atyp_angina – nietypowy ból wynikający ze złego dostarczania tlenu do mięśnia sercowego.
 - Asympt - bezobjawowy
 - non_anginal – ból wynikający z innych przyczyn
- Trestbps: ciśnienie krwi w mmHG w momencie przyjęcia do szpitala – liczba rzeczywista
- Chol: stężenie cholesterolu w surowicy w mg/dl – liczba rzeczywista
- Fbs: czy stężenie cukru we krwi przekracza 120 mg/dl – atrybut nominalny, dwie wartości: t, f.
- Restecg: wynik elektrokardiografii – atrybut nominalny, 3 wartości:
 - Normal – w porządku
 - left_vent_hyper – opis nie dla śmiertelnika
 - st_t_wave_abnormality – opis nie dla śmiertelnika
- Thalach: najwyższe zmierzone tętno – atrybut rzeczywisty.
- Exang: czy ćwiczenia powodują ból w piersi (angina) – atrybut nominalny, 2 wartości: no, yes
- Oldpeak: „ST depression induced by exercise relative to rest”: atrybut rzeczywisty.
- Slope: nachylenie segmentu ST podczas ćwiczeń, atrybut nominalny, 3 wartości:
 - Up,
 - Flat
 - Down



- Ca: liczba głównych żył i tętnic prowadzących do/z serca widocznych podczas fluoroskopii – atrybut rzeczywisty
- Thal: ???, atrybut nominalny, 3 wartości: fixed_defect, normal, reversable_defect
- Num: Diagnoza angiografi, atrybut nominalny, 5 wartości:
 - <50 – mniej niż 50% przewężeń średnicy żył
 - >50_1 – więcej niż 50% przewężeń średnicy żył

- >50_2, >50_3, >50_4 – wartości nie używane

Sonar

Zestaw danych zawiera zbiór pomiarów za pomocą sonaru dotyczących różnych obiektów podwodnych. Z każdym pomiarem związana jest informacja o tym z czego zrobiony obiekt (czy jest to skała czy mina).

- 60 atrybutów liczbowych, rzeczywistych + 1 atrybut decyzyjny nominalny
- 208 pomiarów
- Każdy z 60 atrybutów to ilość energii odpowiadającej określonej częstotliwości dźwięku zebranej (dźwięku odbitego, który „wrócił” do sonaru) podczas pomiaru w jakimś okresie czasu.
- Atrybut decyzyjny ma dwie wartości: rock, mine

Zestaw danych dostępny jest w repozytorium programu Rapid Miner.

mining_build_text

Zestaw danych zawiera wyniki ankiet danych klientom sklepu komputerowego. Poszczególne przykłady zawierają następujące atrybuty:

- CUST_ID - Id klienta, atrybut liczbowy
- CUST_GENDER - Płeć klienta, atrybut nominalny
- AGE - Wiek klienta, atrybut liczbowy
- CUST_MARITAL_STATUS - Stan cywilny, atrybut nominalny
- COUNTRY_NAME - Nazwa kraju klienta, atrybut nominalny
- CUST_INCOME_LEVEL - Poziom dochodów, atrybut nominalny
- EDUCATION – poziom wykształcenia, atrybut nominalny
- OCCUPATION - Aktualnie zatrudniony jako, atrybut nominalny
- HOUSEHOLD_SIZE - Wielkość rodziny, atrybut liczbowy
- YRS_RESIDENCE - Liczba lat zamieszkania pod danym adresem, atrybut liczbowy
- AFFINITY_CARD - Czy posiada kartę zniżkową, atrybut binarny
- BULK_PACK_DISKETTES - Czy zakupił kiedyś paczkę dyskietek, atrybut binarny
- FLAT_PANEL_MONITOR - Czy posiada płaski monitor, atrybut binarny
- HOME_THEATER_PACKAGE - Czy posiada zestaw kina domowego, atrybut binarny
- BOOKKEEPING_APPLICATION - Czy zarządza finansami domowymi za pomocą programu komputerowego, atrybut binarny
- PRINTER_SUPPLIES - Czy zakupił kiedyś materiały do drukarki, atrybut binarny
- Y_BOX_GAMES - Czy posiada konsolę do gier, atrybut binarny
- OS_DOC_SET_KANJI - Czy system operacyjny ma zainstalowane japoński/chiński IME, atrybut binarny
- COMMENTS - Dodatkowy komentarz na ankiecie, atrybut nominalny

Zbiór danych można znaleźć w tabeli MINING_BUILD_TEXT w schemacie DMDATA.

SALES/CUSTOMERS/COUNTRIES/PRODUCTS

Zestaw danych dostępny w schemacie SH przedstawia prostą hurtownię danych firmy handlowej. Schemat składa się z wielu tabel, w tym SALES, CUSTOMERS, COUNTRIES i PRODUCTS. Z każdej tabeli interesuje nas tylko pewien podzbiór atrybutów. Poniżej wymieniono atrybuty wykorzystywane w ćwiczeniach.

COUNTRIES

- COUNTRY_ID – identyfikator kraju, klucz główny
- COUNTRY_NAME - nazwa kraju
- COUNTRY_SUBREGION – nazwa części kontynentu
- COUNTRY_REGION – nazwa kontynentu

PRODUCTS

- PROD_ID – identyfikator produktu
- PROD_NAME – nazwa produktu

CUSTOMERS

- CUST_ID – identyfikator klienta
- CUST_LAST_NAME – nazwisko klienta
- CUST_GENDER – płeć klienta
- CUST_YEAR_OF_BIRTH – rok urodzin
- CUST_MARITAL_STATUS - stan cywilny
- CUST_CITY – miasto zamieszkania
- CUST_STATE_PROVINCE – stan/województwo/dzielnica
- CUST_INCOME_LEVEL - dochody
- COUNTRY_ID – kraj w którym mieszka klient

SALES

- PROD_ID – identyfikator sprzedanego produktu
- CUST_ID – identyfikator klienta
- TIME_ID – moment zawarcia transakcji
- QUANTITY_SOLD – liczba sprzedanych egzemplarzy produktu
- AMOUNT_SOLD – wartość transakcji

Zakupy.arff

Zestaw danych zawierający podzbiór transakcji z zestawu danych SALES, CUSTOMERS, COUNTRIES, PRODUCTS w formacie akceptowalnym przez komponent WEKI który wyszukuje reguły asocjacyjne. W zestawie znajduje się czternaście atrybutów, każdy odpowiada jednemu produktowi. Każdy wiersz odpowiada jednej transakcji. W każdym wierszu, jeżeli dla danego atrybutu zapisane jest 1, to znaczy, że towar został zakupiony, a jeżeli ? to znaczy że tego towaru w ramach transakcji nie kupiono.

Zakupy.binominal.arff

Zawiera te same dane co plik zakupy.arff ale jest w formacie kompatybilnym z komponentem RapidMinera, który zajmuje się wyszukiwaniem reguł asocjacyjnych.

weather.arff/golf

W pliku weather.arff znajduje się zbiór danych opisujących pogodę (outlook, temperature, humidity, windy), oraz atrybut decyzyjny mówiący o tym, czy w taką pogodę dobrze gra się w golfa (play):

- Outlook określający, czy pogoda jest słoneczna, zachmurzona czy deszczowa (attribut nominalny),
- Temperature określający temperaturę (attribut ciągły),
- Humidity określający wilgotność (attribut ciągły),
- Windy określający czy pogoda jest wietrzna czy nie (attribut nominalny),
- Play określający, czy w taką pogodę dobrze gra się w golfa (attribut decyzyjny, nominalny).

weather.nominal.arff

W pliku weather.nominal.arff znajduje się zmodyfikowany zbiór danych zawarty oryginalnie w pliku weather.arff. Modyfikacja obejmuje atrybuty temperature i humidity, które zostały zdyskretyzowane. Zbiór danych zawiera następujące atrybuty:

- Outlook określający, czy pogoda jest słoneczna, zachmurzona czy deszczowa (attribut nominalny),
- Temperature określający czy temperatura jest chłodna (cool), przeciętna (mild) czy wysoka (hot) (attribut nominalny),
- Humidity określający czy wilgotność jest wysoka, czy niska (attribut nominalny),
- Windy określający czy pogoda jest wietrzna czy nie (attribut nominalny),
- Play określający, czy w taką pogodę dobrze gra się w golfa (attribut decyzyjny, nominalny)

Iris.arff

Iris.arff jest zbiorem danych opisującym zmierzone wartości parametrów kwiatów trzech gatunków Irysów: Iris Setosa, Iris Versicolour i Iris Virginica. Każdy kwiat opisany jest pięcioma atrybutami:

- Sepal length – długość w centymetrach listka kielicha kwiatowego (attribut ciągły),
- Sepal width – szerokość w centymetrach listka kielicha kwiatowego (attribut ciągły),
- Petal length – długość w centymetrach płatk kwiatu (attribut ciągły),
- Petal width – szerokość w centymetrach płatk kwiatu (attribut ciągły),
- Class – gatunek irysa (Iris Setosa, Iris Versicolour lub Iris Virginica) (attribut decyzyjny, nominalny)

Na każdą klasę przypada 50 opisów kwiatów. Niniejszy zbiór danych jest bardzo popularny w literaturze dotyczącej uczenia maszynowego. Jedna z klas jest liniowo separowana od dwóch pozostałych.

car.arff

W pliku car.arff znajduje się opis pewnego zbioru samochodów, oraz związaną z nimi decyzją, czy taki samochód jest dla nas akceptowalny czy nie (czy byśmy go kupili czy nie). Każdy samochód opisany jest następującymi atrybutami nominalnymi:

- buying – cena kupna samochodu (niska, średnia, wysoka, bardzo wysoka)

- maint – koszt utrzymania samochodu (niska, średnia, wysoka, bardzo wysoka)
- doors – liczba drzwi (2,3,4, 5 i więcej)
- persons – liczba osób wchodząca do samochodu (2,4 lub więcej)
- lug_boot – wielkość bagażnika (mała, średnia duża)
- safety – bezpieczeństwo jazdy samochodem (niska, średnia wysoka)
- class – czy samochód jest dla nas akceptowalny (nieakceptowalny, akceptowalny, dobry, bardzo dobry) (atrybut decyzyjny)

zoo.arff/ZOO

W pliku zoo.arff znajduje się opis pewnej liczby zwierząt. Ten sam zbiór danych można znaleźć w tabeli ZOO, w schemacie DMDATA. Zwierzęta opisane są następującymi atrybutami nominalnymi:

- hair – czy zwierzę ma włosy
- feathers – czy zwierzę ma pióra
- eggs – czy zwierzę znosi jaja
- milk – czy karmi młode mlekiem
- airborne – czy zwierze potrafi latać
- aquatic – czy zwierze żyje w wodzie
- predator – czy jest drapieżnikiem
- toothed – czy posiada zęby
- backbone – czy posiada kręgosłup
- breathes – czy oddycha
- venomous – czy jest jadowite
- fins – czy ma płetwy
- legs – liczba nóg
- tail – czy ma ogon
- domestic – czy zostało udomowione
- type – typ (mammal, bird, reptile, fish, amphibian, insect, invertebrate) (atrybut decyzyjny)

spam.arff

W pliku spam.arff znajduje się zbiór opisów 4600 maili, którym przydzielono etykietę, czy stanowią SPAM, czy nie. Każdy opis maila składa się z 58 atrybutów, z czego 57 jest atrybutami ciągłymi, a jeden (decyzyjny – „class”) jest nominalny i wynosi 1, jeżeli mail stanowi SPAM i 0 jeżeli mail nie jest traktowany jako SPAM. Pozostałe 57 atrybutów określa częstotliwość występowania wyrazów charakterystycznych dla spamu, znaków przestankowych oraz długość ciągów dużych liter.

flags.arff

Plik flags.arff zawiera ok. 120 zbiorów elementów opisanych poniżej,

- Bars – flaga zawiera poziome paski;
- Stripes – flaga zawiera pionowe paski;
- Red – flaga zawiera kolor czerwony;
- Green – flaga zawiera kolor zielony;
- Blue – flaga zawiera kolor niebieski;
- Gold – flaga zawiera kolor żółty lub złoty;
- White – flaga zawiera kolor biały;

- Black – flaga zawiera kolor czarny;
- Orange – flaga zawiera kolor pomarańczowy;
- Circles – flaga zawiera kółka;
- Crosses – flaga zawiera krzyże „proste”;
- Saltires – flaga zawiera krzyże „ukośne”;
- Quarters – flaga podzielona jest na sekcje;
- Sunstars – flaga zawiera symbole słońca i/lub gwiazd;
- Crescent – flaga zawiera półksiężyc;
- Triangle – flaga zawiera trójkąty;
- Icon – flaga zawiera rysunek obiektu nieożywionego;
- Animate – flaga zawiera rysunek obiektu ożywionego;
- Text – flaga zawiera tekst

Każdy zbiór opisuje jedną z flag krajów na świecie.

2circles.arff

Zestaw danych zawierający dwa atrybuty numeryczne X i Y reprezentujące współrzędne punktów w dwuwymiarowej przestrzeni. Chmury punktów układają się w dwie figury: koło i „okrąg” o wspólnym środku.

birth.arff

Zestaw danych zawierający trzy atrybuty:

- Country – nazwa kraju (attribut nominalny)
- BirthRate – współczynnik urodzeń w kraju (attribut ciągły);
- DeathRate – współczynnik śmiertelności w kraju (attribut ciągły);

milk.arff

Zestaw zawiera skład mleka 25 różnych gatunków ssaków. Zawiera 6 atrybutów:

- Species – nazwa gatunku (attribut nominalny)
- Water – procent zawartości wody (attribut ciągły);
- Protein – procent zawartości białka (attribut ciągły);
- Fat – procent zawartości tłuszczu (attribut ciągły);
- Lactose – procent zawartości laktozy (attribut ciągły)
- Ash – procent masy jaki pozostaje po spaleniu (attribut ciągły)

AUTOS

TBD

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all

autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

Note: Several of the attributes in the database could be used as a "class" attribute.

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 to 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcvt, l, ohc, ohcvt, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

MOVIES/BAD_MOVIES/RATINGS

TBD

TITANIC

TBD

VOTES

TBD

CLOUDS

TBD

EMP

TBD